A joint model for author's characteristics and collaboration pattern in bibliometric networks: a Bayesian approach

Stefano Nasini*   Víctor Martínez-de-Albéniz†   Tahereh Dehdarirad‡

August 25, 2015

Abstract

Demographic and behavioral characteristics of journal authors are important indicators of *homophily* in co-authorship networks. In the presence of correlations between adjacent nodes (assortative mixing), combining the estimation of the individual characteristics and the network structure results in a well-fitting model, which is capable to provide a deep understanding of the linkage between individual and social properties. This paper aims to propose a novel probabilistic model for the joint distribution of nodal properties (authors' demographic and behavioral characteristics) and network structure (co-authorship connections), based on the nodal similarity effect. A Bayesian approach is used to estimate the model parameters, providing insights about the probabilistic properties of the observed data set. After a detailed analysis of the proposed statistical methodology, we illustrate our approach with an empirical analysis of co-authorship of 1007 journal articles indexed in the ISI Web of Science database in the field of neuroscience between 2009 and 2013.

Key words: Bibliometrics, Social networks, Co-authorship networks, Nodal similarities, Bayesian inference, MCMC.

1 Introduction

The increasing specialization of scientific research, the interdisciplinary character of most projects, and the increased funding of cross-institution initiatives have made scientists take part in collaboration networks (Teixeira da Silva 2011, Haeussler and Sauermann 2013). Co-authorship networks represent a widely studied class of collaboration networks. They have been extensively studied using different statistical approaches (Newman 2004, 2003), with the purpose of identifying the structure of scientific partnerships and the role played by the individual researchers.

---

*Corresponding author: IESE Business School, University of Navarra, Barcelona, Spain, SNasini@iese.edu

†IESE Business School, University of Navarra, Barcelona, Spain, valbeniz@iese.edu. V. Martínez-de-Albéniz's research was supported in part by the European Research Council - ref. ERC-2011-StG 283300-REACTOPS and by the Spanish Ministry of Economics and Competitiveness (Ministerio de Economía y Competitividad) - ref. ECO2011-29536.

‡Department of Library and Information Science, University of Barcelona, Barcelona, Spain, tdehdari@gmail.com

The majority of methodological contributions focus on modelling the structure of scientific co-authorship, based on the projection of a two-mode network (author–paper network) into a one-mode structure of co-authorship (author–author network), where links represent co-authors, i.e., authors sharing common papers, as described by Leydesdorff and Wagner (2008). We use a similar approach here. Our main contribution is to combine individual and social properties of journal authors, by internalizing the effect of authors' similarities in their patterns of connection. It provides an insight into the level of *homophily* in co-authorship networks, in terms of specific socio-demographic characteristics (Newman 2003), while accounting for relevant network features based on observed nodal properties.

The notation we use denotes $\mathcal{V}$ the set of N authors and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ their known structure of connections; $\mathcal{K}$ denotes a set of K of categorical properties (in our application, $\mathcal{K} = \{genders, \, nationalities\}$) defined for each author in $\mathcal{V}$. The nodal similarities are assumed to reflect the overlap of authors' categorical statuses, with respect to the properties in $\mathcal{K}$.

An exponential random model is proposed to internalize the effect of nodal similarities on the joint distribution of network and authors' properties (Caimo and Friel 2011, Robins et al. 2007). Exponential families possess good properties that typically simplify the statistical inference of model parameters. As we explain in Section 3, the inclusion of nodal similarities as sufficient statistics for the exponential random model entails the impossibility of a complete characterization of the probability distribution, due to the intractability of the normalizing constant. This represents one of the strongest barriers to the numerical-optimization of the likelihood-function and legitimates the use of simulation-based approaches – such as the Monte Carlo maximum likelihood of Geyer and Thompson (1992) and pseudo-likelihood estimation of Strauss and Ikeda (1990).

As suggested by Caimo and Friel 2011, this drawback can be overcome by embedding the defined model into a Bayesian estimation framework, which reformulate the estimation problem based on the ability of simulating from the posterior distribution. We build on Murray et al. 2006, which proposed a MCMC method to simulate from this class of distributions, allowing a flexible estimation of the effect of nodal similarity – which is the main scope of this paper. Our approach is able to generate the following insights:

- *we can estimate author' collaborations based on their demographic and behavioral characteristics;*

- *we can estimate author' demographic and behavioral characteristics based on their pattern of connections.*

In other words, our model connects nodal properties with network structure, so one can be used to predict the other. We illustrate our method through the analysis of co-authorship of over a thousand journal articles between 2009 and 2013 in the neuroscience research community.

Section 2 introduces and describes the co-authorship data set, along with the relevant network statistics we aim to control in a probabilistic model. Section 3 provides a detailed description of the proposed exponential random model for this type of data set and embeds such model in a general Bayesian framework. Section 4 takes into account the algorithmic aspects of the estimation of the model parameters. The numerical results are presented in Section 5. Section 6 concludes.

2 Presenting the co-authorship data set

The data set is composed of the scientific publications indexed in the Web of Science (WOS) database between 2009 and 2013 in the field of Neuroscience. 153,182 research papers were retrieved in the first step. Then, we conducted stratified random sampling. The sample size was determined with a 3% sampling error and 95% of level of confidence. Table 1 shows the total number of publications and the stratified sample size per year in the studied field.

Year	# publications (%)	Stratified sample size
2009	28,819 (18.81%)	199
2010	30,154(19.69%)	208
2011	31,030 (20.26%)	214
2012	31,265 (20.41%)	218
2013	31,914 (20.83%)	221
Total	153,182	1,060

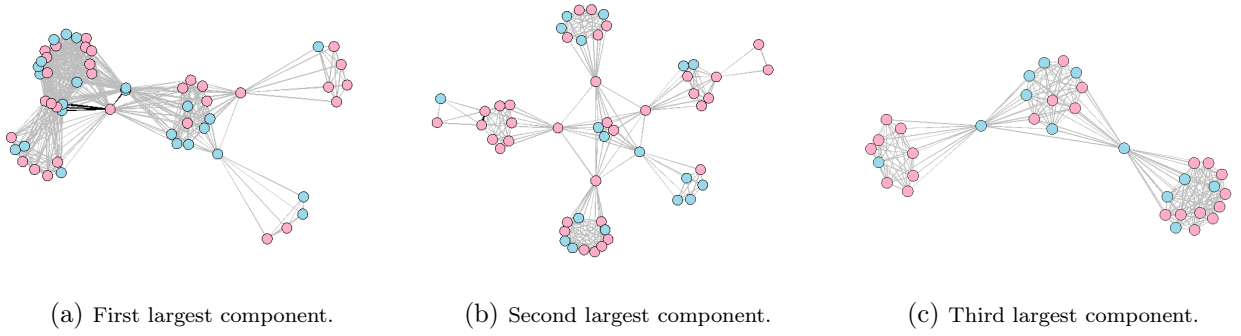Table 1: The total number of publications and the stratified sample size, 2009-2013.

(a) First largest component.   (b) Second largest component.   (c) Third largest component.

Figure 1: The three largest components with nodal genders (blue for men, pink for women).

(a) First largest component.   (b) Second largest component.   (c) Third largest component.
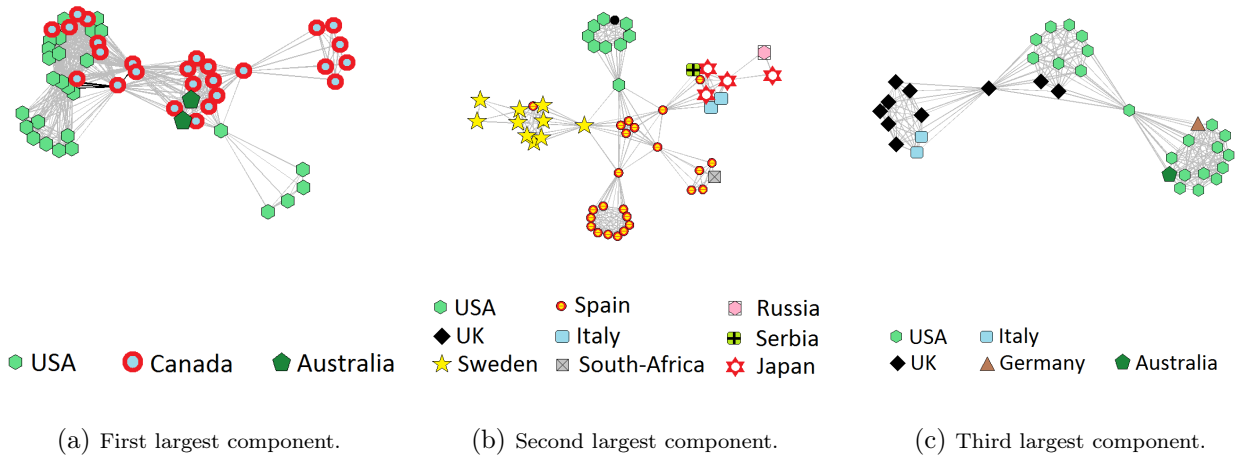
Figure 2: The three largest components with nodal nationalities.

Finally, after eliminating those papers whose gender was unclear (53 of them, 5%), our data set comprised 1,007 (95%) of 1,060 papers. These $1,007$ papers were used as our data set for further analysis, corresponding to $5,385$ authors. Each author was assigned to a nationality based on the authors' country of affiliation. Thus, for each author, two demographic characteristics were collected: gender and nationality.

A network structure of scientific collaboration between authors was then generated by connecting those authors whose names jointly appear in one or more of the $1,007$ articles. The resulting network comprised 207 disconnected components and the three largest had respectively size of 55, 53 and 35, as shown in Figures 1 and 2, associated with the corresponding nodal gender and nationalities.

3 Model definition and specification

This section proposes an exponential random model which internalizes the structure of dependencies of individual characteristics and network structure. Consider a categorical variable $\mathbf{Y}_k$, with m_k categories defined on a set of N individuals, and its representation in term of an $N \times m_k$ binary matrix $\mathbf{y}_k \in \{0,1\}^{m_k n}$. Let $\mathcal{Y}_k$ be the set of all possible realizations of $\mathbf{Y}_k$. Similarly, let $\mathbf{Z}$ be the adjacency matrix of a random network with N nodes and $\mathcal{Z} \subseteq \{0,1\}^{N \times N}$ the set of its possible realizations. The sample space under consideration can be defined as $\mathcal{X} = \mathcal{Z} \times \mathcal{Y}_1 \times \ldots \times \mathcal{Y}_K \subseteq \{0,1\}^{N \times N + \sum_{k=1}^{K} m_k}$ – the set of network structures among N individuals, taking K categorical properties, as illustrated in Figure 3.
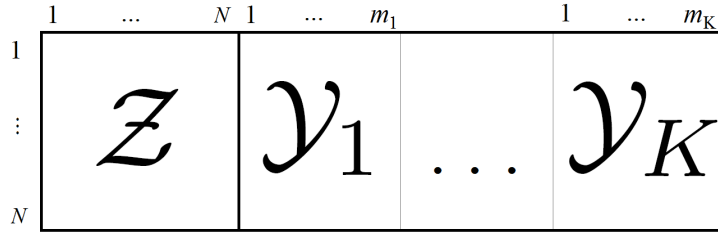

Figure 3: Sample space.

Let $\mathbf{X}$ be a random matrix, taking values in $\mathcal{X}$ and $\mathbf{x}$ a possible realization. In the exponential family of distributions the conditional probability of $\mathbf{x} \in \mathcal{X}$ takes the following form: $P(\mathbf{x} \mid \theta) \propto \exp(T(\mathbf{x})^{\mathrm{T}}\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of *natural parameters* of the distribution, which can usually take any value in the reals; and $T(\mathbf{x})$ is a vector of sufficient statistics.

The exponential random model developed here internalizes the nodal similarity effect by including the matching in the categorical properties of each pair of nodes as a sufficient statistic of the exponential distribution:

$$P(\mathbf{x} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto \begin{cases} \exp\left[\boldsymbol{\alpha}^{\mathrm{T}} A(\mathbf{y}) + \boldsymbol{\beta}^{\mathrm{T}} B(\mathbf{z}) + \boldsymbol{\gamma}^{\mathrm{T}} G(\mathbf{y}, \mathbf{z})\right] & \text{if } \mathbf{x} \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $\mathbf{y}$ is a matrix whose (h_k, r) component represents the dummy indicator of whether an individual

r has value h_k in the kth categorical variable, for $h_k = 1 \ldots m_k$, $r \in \mathcal{V}$; $\mathbf{z}$ is the adjacency matrix whose (r, s) component represents the binary indicator of the existence of a connection between $r \in \mathcal{V}$ and $s \in \mathcal{V}$. $A(\mathbf{y})$ is a vector of sufficient statistics which accounts for combinatorial properties of the categorical variables $\mathbf{y}$ (such as the number of nodes per each level of each categorical variable, number of associated categories, etc.) only, and $\boldsymbol{\alpha}$ is the corresponding vector parameter. Similarly, $B(\mathbf{z})$ is a vector of sufficient statistics which accounts for combinatorial properties of the network structure $\mathbf{z}$ (such as the clustering coefficient, the assortativity coefficient, the average path length, etc.), but independent of nodal exogenous properties; and $\boldsymbol{\beta}$ is the corresponding vector parameter. The interaction between nodal characteristics and connections variables is internalized into the model by the sufficient statistics $G(\mathbf{y}, \mathbf{z})$; $\boldsymbol{\gamma}$ is the parameter vector associated with these interactions.

The specification of the sample space $\mathcal{X}$ can incorporate both network and nodal properties, in accordance with our modeling assumptions and our need to control specified combinatorial properties (Castro and Nasini 2015). In other words, $P(\mathbf{x} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = 0$ if $\mathbf{x}$ does not satisfy a set of feasibility constraints. As illustration, three possible sample spaces are shown in (3) by exogenously fixing the degree sequence, the number of edges and the size of each categorical level. They are specified in term of the solution sets of systems of linear constraints. Note that intersections of these sets give rise to hybrid sample spaces with complex combinatorial structures.

$$
\text{fixed degree sequence} \quad \left|
\begin{array}{ll}
\sum_{h_k=1}^{m_k} y_{h_k r} = 1 & k = 1 \ldots K,\, r = 1 \ldots n \\
\sum_{r=1}^{n} z_{rs} = d_s & r \in \mathcal{V}
\end{array}
\right.
$$

$$
\text{fixed number of edges} \quad \left|
\begin{array}{ll}
\sum_{h_k=1}^{m_k} y_{h_k r} = 1 & k = 1 \ldots K,\, r = 1 \ldots n \\
\sum_{(r,\,s) \in \mathcal{V} \times \mathcal{V}} z_{rs} = d &
\end{array}
\right.
$$

$$
\text{fixed categorical levels} \quad \left|
\begin{array}{ll}
\sum_{h_k=1}^{m_k} y_{h_k r} = 1 & k = 1 \ldots K,\, r = 1 \ldots n \\
\sum_{r=1}^{m} y_{h_k r} = f_k & k = 1 \ldots K,
\end{array}
\right.
$$

Classical inferential methods for the model parameters of (1) are encumbered by the intractability of the normalizing constant, which makes the numerical optimization of the likelihood function very challenging. The next section describes an MCMC algorithm to simulate from the Bayesian posterior distribution of the parameter.

4 Estimation method

As noted by Murray et al. (2006) and by Caimo and Friel (2011), the intractability of the normalizing constants of most random network models entails a "double intractability" of the posterior distribution when the model is embedded in a Bayesian framework. This is also true for model (1). MCMC algorithms are often used to draw samples from distributions with intractable normalization constants. However, they do not apply to a doubly-intractable constant.

Consider the kernel of the probability function (1) and let $\mathbf{x}^{(0)} \in \mathcal{X}$ be the observed data set –the co-authorship network structure $\mathbf{z}^{(0)}$, the nodal genders $\mathbf{y}_1^{(0)}$, the nodal nationalities $\mathbf{y}_2^{(0)}$. Given a prior distribution $\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$, apply the Bayes rule:

$$P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{x}^{(0)}) = \frac{P(\mathbf{x}^{(0)} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}{\displaystyle\int_{\alpha, \beta, \gamma} P(\mathbf{x}^{(0)} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \, d\boldsymbol{\alpha} \, d\boldsymbol{\beta} \, d\boldsymbol{\gamma}}$$

Since both $P(\mathbf{x}^{(0)} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$ and $P(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{x}^{(0)})$ can only be specified under proportionality conditions, Murray et al. (2006) proposed a MCMC approach which overcomes the drawback to a large extent, based on the simulation of the joint distribution of the parameter and the sample spaces, conditioned to the observed data set $\mathbf{x}_0$, that is to say, $P(\mathbf{x}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{x}_0)$. We follow the same approach. Our application of the Metropolis-Hastings method (Bolstad 2009) to simulate from such distribution is summarized in Algorithm 1.

Algorithm 1 Exchange algorithm of Murray et al. (2006).

1: Initialize $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
2: **repeat**
3:     Draw $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ from $h(. \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
4:     Draw $\mathbf{x}'$ from $P(. \mid \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$
5:     Accept $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ with probability $\min\left\{1, \dfrac{P(\mathbf{x}' \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})P(\mathbf{x}^{(0)} \mid \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')\pi(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')}{P(\mathbf{x}^{(0)} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})P(\mathbf{x}' \mid \boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')\pi(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})}\right\}$
6:     Update $(\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma})$
7: **until** Convergence

The distribution $h()$ is used to simulate candidate points from the posterior and it is here assumed to be symmetric. Note that in step 3 of Algorithm 1 a new value of the parameters $(\boldsymbol{\alpha}', \boldsymbol{\beta}', \boldsymbol{\gamma}')$ is randomly proposed and in step 4 a sample from $\mathcal{X}$ is simulated with probability given in (1). Clearly, this is a computationally intensive procedure.

5 Numerical results and analysis

In this section the described probabilistic model is applied to the second largest component of the co-authorship network in the neuroscience community, as shown in figures 1 and 2, where $n = 54$ (nodes), $K = 2$ (categorical properties), $m_1 = 2$ (genders), $m_2 = 9$ (nationalities). The sample space is defined as the Cartesian product between the set of n node undirected networks with fixed number

of edges $d = 234$ and the set of all possible realization of 2 categorical variables with 2 and 9 levels. The following specification of model (1) is taken into account in this section

$$P(\mathbf{x}) \propto \begin{cases} \exp\left[\sum_{k \in \mathcal{K}} \sum_{h_k=1}^{m_k} \alpha_{k,h_k} \left(\sum_{r \in \mathcal{V}} y_{h_k r}\right) + \sum_{k \in \mathcal{K}} \gamma_k \sum_{(r,s) \in \mathcal{V} \times \mathcal{V}} z_{rs} \left(\sum_{h_k=1}^{m_k} y_{h_k r} y_{h_k s}\right)\right] & \text{if } \mathbf{x} \in \mathcal{X} \\ 0 & \text{otherwise} \end{cases}$$
(2)

The vector of sufficient statistics thus contains the total amount of nodes for each property and each level, and the association between edges and two nodal properties The corresponding natural parameters are α_{k,h_k} and γ_k.

It is important to keep in mind which interpretation should be given to the estimated natural parameters α_{k,h_k} and γ_k, for $h_k = 1 \ldots m_k$, $k \in \mathcal{K}$. In the case of uniform distribution within the sample space $\mathcal{X}$, we should have $\alpha_{k,h_k} = 0$; hence deviations in $\alpha_{k,h_k} \neq 0$ can be interpreted as incorporating different proportions of the properties within the nodal population (e.g., nationalities not being evenly present in the sample). Moreover, if nodal and network properties are independent, then $\gamma_k = 0$: the association between connections and nodal similarities has no effect on the probability of observing a given configuration in $\mathcal{X}$. Any non-zero value of the natural parameters entails a deviation from such independence. Specifically, if $\gamma_k = 0$, (2) is reduced to the product between the probability mass functions of K multinomial random variables $\text{Multinom}(e^{\alpha_{k,1}} \ldots e^{\alpha_{k,m_k}})$ and the Erdos-Reniy random Graph model with fixed number of edges d. In the numerical results we set $\gamma_k = 0$ as a null model for the computational comparison.

The contour plot in Figure 4 show the estimated marginal posterior of (γ_1, γ_2), corresponding to the *gender* and the *nationality* effect, for the second largest component of the co-authorship data set in Section 2. These results have been obtained by 3 chains with 30,000 MCMC iterations.
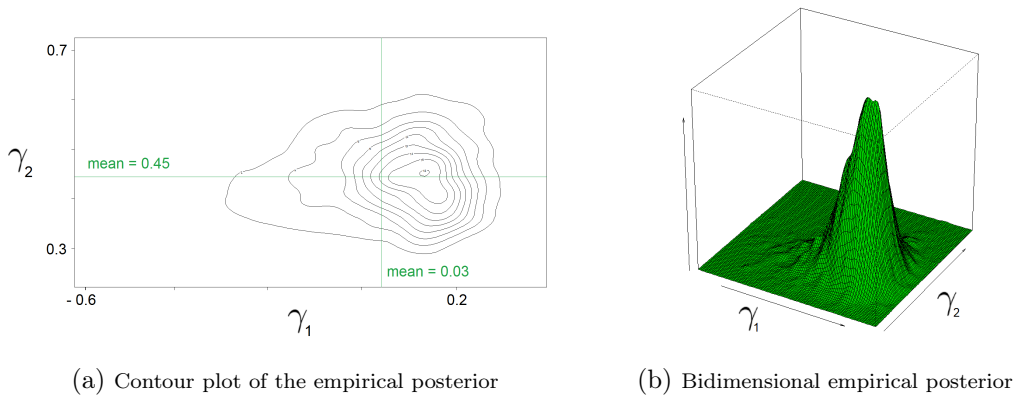

(a) Contour plot of the empirical posterior   (b) Bidimensional empirical posterior

Figure 4: Marginal posterior of (γ_1, γ_2), corresponding to the second largest component of the co-authorship data set.

Figure 4 reports a positive expected effect of the nationality (0.45) and a negligible expected effect of the gender (0.03). It can be noted a much larger variability on the gender effect, suggesting a lack of information about the parameter γ_1.

After estimating the model parameters $(\boldsymbol{\alpha}, \boldsymbol{\gamma})$, a sample of 10,000 elements from $\mathcal{X}$ has been

simulated and the nodal similarities have been computed, both for the estimated model and the null model $(\boldsymbol{\alpha}, \mathbf{0})$. Figure 5 shows the observed and expected nationality similarities for each of the $54 \times 53/2$ pairs of authors. It can be noted that the null model seems to be severely unable to capture the pairwise similarities, whereas the expected values of $\sum_{h_k=1}^{m_k} y_{h_k r} y_{h_k s}$ under the estimated model in (2) substantially resemble the observed similarities.
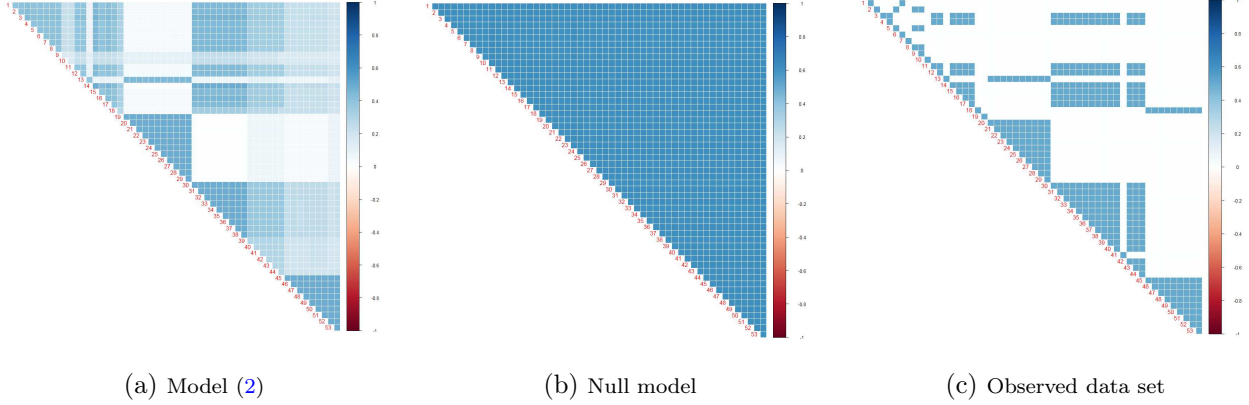

(a) Model (2)   (b) Null model   (c) Observed data set

Figure 5: The values of the nationality similarities $\sum_{h_k=1}^{9} y_{h_k r} y_{h_k s}$ for the $54 \times 53/2$ pairs of nodes .

Tables 2 and 3 reports the expected proportions of edges associated to different combinations of genders and nationalities respectively. The estimated model seems to be able to effectively capture the assortative mixing of the data, i.e., the association between nodal properties and connections, along with the total amount of each individual categories and network collaboration density. This was indeed our initial intention, when a joint model for author's characteristics and collaboration pattern was introduced in Section 3.

	male	female	total
male	0.40 (0.48)	0.46 (0.40)	0.86 (0.88)
female	$--$	0.14 (0.12)	
total		0.60 (0.52)	

Table 2: Expected proportions of edged for each combination of genders. The observed proportions are reported within parenthesis.

	Italy	USA	Spain	Sweden	S. Africa	Japan	Serbia	Russia	UK
Italy	0.004 (0.004)	0.004 (0.000)	0.014 (0.017)	0.001 (0.025)	0.000 (0.008)	0.007 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
USA		0.260 (0.235)	0.120 (0.031)	0.012 (0.000)	0.006 (0.000)	0.060 (0.004)	0.005 (0.000)	0.000 (0.000)	0.005 (0.023)
Spain			0.267 (0.329)	0.020 (0.024)	0.010 (0.009)	0.111 (0.064)	0.010 (0.201)	0.000 (0.000)	0.010 (0.000)
Sweden				0.001 (0.017)	0.001 (0.006)	0.010 (0.000)	0.000 (0.001)	0.000 (0.008)	0.000 (0.000)
S.Africa					0.001 (0.006)	0.002 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Japan						0.085 (0.140)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Serbia							0.000 (0.000)	0.000 (0.000)	0.000 (0.000)
Russia								0.000 (0.000)	0.000 (0.000)
UK									0.000 (0.000)

Table 3: Expected proportions of edged for each combination of nationalities. The observed proportions are reported within parenthesis.

In Figure 6 a graphical comparison between the expected (dark-grey) and the observed (light-grey) degree sequences is provided.
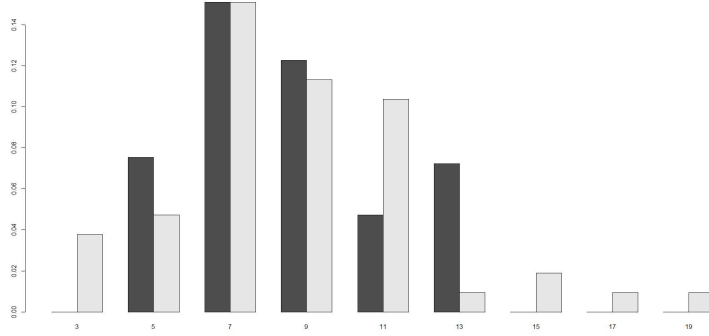

Figure 6: In light-grey the observed degree sequence. In dark-grey the expected degree sequence.

To summarize, for the second largest component of the co-authorship data set, the estimated results confirm the null effect of the gender similarity on the author's connections, along with a positive effect of their nationalities. The estimated model seems to properly fit the empirical observation, as suggested by table 2 and 3. The expected degree sequence also resembles the observed one, suggesting the ability of the model to capture both individual and structural properties of the co-authorship data.

6 Conclusion

This paper presented an exponential random model for author's characteristics and collaboration pattern in bibliometric networks, which allowed to combine the analysis of multivariate data set with the one of the assortative pattern of nodal similarities in networks. We proposed a Bayesian estimation framework and a specialized MCMC algorithm to simulate from a "doubly intractable" posterior distribution. We showed a strong capability of the model to account for relevant network features (the degree sequence) based on the observed nodal properties, providing a deep understanding of the linkage between individual and social properties and a substantial insight into the level of *homophily*

in co-authorship networks.

Our results suggest several lines of work for future research. We could compare the model fit for different specifications of the sample space $\mathcal{X}$. We could also study the inclusion of further nodal properties, such as age, principal keywords or number of received citations.

References

Bolstad, W. M. 2009. *Markov chain monte carlo sampling from posterior*, 127–157. John Wiley and Sons, Inc.

Caimo, A., and N. Friel. 2011. Bayesian inference for exponential random graph models. *Social Networks* 33 (1): 41–55.

Castro, J., and S. Nasini. 2015. Mathematical programming approaches for classes of random network problems. *European Journal of Operational Research* 2 (4): 317 – 325.

Geyer, C. J., and E. A. Thompson. 1992. Constrained Monte Carlo Maximum Likelihood for Dependent Data. *Journal of the Royal Statistical Society. Series B (Methodological)* 54 (3): pp. 657–699.

Haeussler, C., and H. Sauermann. 2013. Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship. *Research Policy* 42 (3): 688 – 703.

Leydesdorff, L., and C. S. Wagner. 2008. International collaboration in science and the formation of a core group. *Journal of Informetrics* 2 (4): 317 – 325.

Murray, I., Z. Ghahramani, and D. J. C. MacKay. 2006. MCMC for doubly-intractable distributions. In *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, 359–366. AUAI Press.

Newman, M. E. J. 2003, February. Mixing patterns in networks. *Phys. Rev. E* 67 (2): 026126.

Newman, M. E. J. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences* 101 (suppl 1): 5200–5205.

Robins, G., T. Snijders, P. Wang, M. Handcock, and P. Pattison. 2007. Recent developments in exponential random graph (p*) models for social networks. *Social Networks* 29 (2): 192 – 215. Special Section: Advances in Exponential Random Graph (p*) Models.

Strauss, D., and M. Ikeda. 1990. Pseudolikelihood Estimation for Social Networks. *Journal of the American Statistical Association* 85 (409): pp. 204–212.

Teixeira da Silva, J. A. 2011. The ethics of collaborative authorship. *EMBO reports* 12 (9): 889–893.